Surgeon-performed ultrasonography

Collecting validity evidence for assessment of abdominal and head & neck ultrasonography skills

Tobias Todsen

This review has been accepted as a thesis together with three previously published papers by University of Copenhagen 26th of September 2016 and defended on 1st of December 2016.

Tutors:
Lars Konge, Morten Lind Jensen and Charlotte Ringsted.

Official opponents:
Teodor Grantcharov, Dorte Guldbrand Nielsen and Lars Bo Svendsen.

Correspondence: Copenhagen Academy for Medical Education and Simulation, University of Copenhagen and the Capital Region of Denmark, Blegdamsvej 9, 2100 Copenhagen East

E-mail: tobiastodsen@gmail.com

Dan Med J 2017;64(11):B5421

THE 3 ORIGINAL PAPERS ARE

1. Todsen T, Tolsgaard MG, Olsen BH, Henriksen BM, Hillingsø JG, Konge L, Jensen ML, Ringsted C. Reliable and Valid Assessment of Point-of-Care Ultrasonography. Ann Surg. 2015 Feb;261(2):309-15

2. Todsen T, Jensen ML, Tolsgaard MG, Olsen BH, Henriksen BM, Hillingsø JG, Konge L, Ringsted C. Transfer from point-of-care Ultrasonography training to diagnostic performance on patients – a randomized controlled trial. Am J Surg. 2016 Jan;211(1):40-5

3. Todsen T, Melchiors J, Charabi B, Henriksen B, Ringsted C, Konge L, von Buchwald C. Competency-based assessment in surgeon-performed head and neck ultrasonography: A validity study. Laryngoscope. 2017 Sep. [Epub ahead of print]

BACKGROUND

Ultrasonography as diagnostic imaging technique

Ultrasonography (US) has been used for diagnostic imaging since the development of the B-mode scanner in the early 1950s (4). It is considered a safe imaging modality in contrast to the traditional diagnostic modalities based on ionizing radiation (5). In brief, a transducer produces high-frequency sound waves, which penetrate the relevant part of the body and are reflected back to the transducer according to the echogenicity of the tissue. This information is then converted by the US machine into a diagnostic image, typically a two-dimensional gray-scale image as presented in Figure 1.

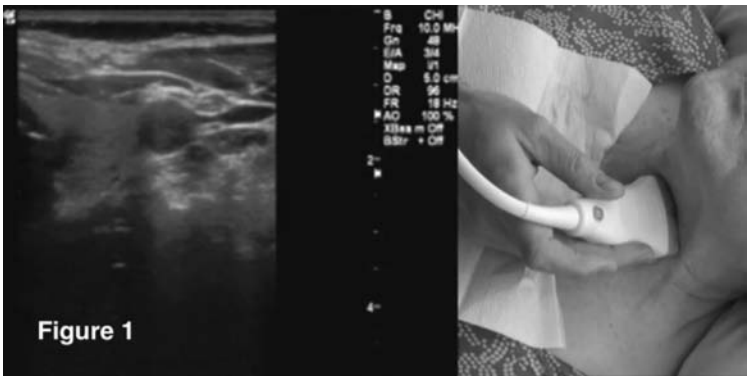

Figure 1. Demonstration of a head & neck US examination.

Like other imaging modalities, US is usually provided by the radiology department based on a request from the physician being responsible for the patient treatment (6). Here US is performed by a radiologist or by a radiographer who performs the US examination and stores it for review by the radiologist (7). Based on the US examination the radiologist writes a report that is used by the physician to decide the further medical management of the patient. US is a dynamic real-time image modality that – unlike static imaging conducted with standardized scanning protocols – can be difficult for other than the US operator to reproduce. Due to the development of low-cost portable high-resolution US ma- chines, US is increasingly being performed bedside by the physi- cian as an adjunct to the physical examination (8,9). As opposed to the comprehensive examinations of the complete organ sys- tem performed at the radiology department on request (10), point-of-care US often answers a focused clinical question, e.g. does the patient have free fluid in the peritoneum? (9). Other terms such as “clinical US,” “bedside US,” and “focused US,” are used in the literature to describe the same concept (11). Especial- ly surgeons have advantage of their unique anatomy knowledge (both from a visually and tactile perspective from surgical experi- ence within the area) when they perform and interpret US (12). Surgeon-performed US was introduced in trauma surgery to detect abdominal and pericardial bleeding in patients with blunt traumas (13). Here US is used in the decision to determine if the arrived patient should be transferred directly to the operating room or should be referred to a computed tomography (CT) scan (14). Further, the portable US machines also made it possible for surgeons and emergency physicians to diagnose patients with

acute abdominal pain in the emergency room (15). Because radiologists are rarely available around the clock, US is increasingly being used in the emergency setting to triage and to decide on the initial patient management (16). The point-of-care US of the abdomen do therefore not replace radiologist-performed abdominal US, but is rather used as an extension to the physical examination of the acute patient. US is therefore relevant to most surgeons because it allows to quickly assess patients' anatomical and physiological characteristics. However, the deep structures and lesions acoustically shadowed by bone or air is difficult to visualize with US, which can limit the use of US of abdomen compared to CT. In contrast, the superficial structures of the head & neck are optimal for US–providing images with higher resolution of lymph nodes, thyroid, and the salivary glands than CT (17-20). US is therefore the first-line imaging modality for patients referred with neck masses (21) and is essential in TNM staging of thyroid cancer (22,23). The portable high-resolution US machines have made office-based US possible and head & neck surgeons are increasingly performing US in their out-patient clinics (24,25). It can be used in the preoperative planning to characterize neck lesions in detail and explore its anatomic relation to adjoining structures in real time (26). US is also recommended as guidance of fine-needle aspiration in the diagnostic work-up of neck masses, and surgeon-performed US can save the patient for an additional appointment at a radiology department (27). So in contrast to the surgeon-performed US in the emergency setting, the office-based US may increasingly replace the radiologist-performed US (12,28).

However, US is a very user-dependent image modality and competence of the operator are needed in order to ensure high diagnostic accuracy. This thesis will therefore explore how we can ensure competence in surgeon-performed US in both an emergency setting and in office-based setting for head & neck lesions.

Certification in ultrasonography

The current certification in US follow a traditional model using the number of US examinations performed over a period of time as indicator of achieved competence (29,30). According to the guidelines from the US societies – European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) and The American Institute of Ultrasound in Medicine (AIUM) – formal US training and a fixed number of supervised US examination are needed to be deemed competent in US (31,32). The current certification demands from American and European US societies require between 300-500 supervised abdominal US examinations (33,34) and 150 supervised head & neck US examinations (35). The EFSUMB has no official recommendation for head & neck US training but in Germany the Deutsche Gesellschaft für Ultraschall in der Medizin (DEGUM) requires 400 supervised head & neck US examinations before certification for independent practice of US (36). In Denmark residents in general surgery and otolaryngology already have requirements of formal 'hands-on' training in point-of-care US in order to receive board certification (37,38). However, the residents do not have any requirement of number of US examinations needed in order to complete their training. In general, there is a lack of consensus about the educational requirements to surgeon-performed US, present being based on expert consensus rather than emperical research. However, a shift towards competence-based surgical education has emerged over the recent years and emphasized the need for objective and reliable skills assessments (39-41). We know that physicians have different learning curves in US why performance-based assessment is more reliable than using a fixed number of completed examinations (42). For this purpose we need an assessment tool with validity evidence supporting the assessment of US skills. In the next paragraph I will analyze the skills needed by the US operator from an educational perspective and thereafter discuss relevant theories used in skills assessment. In the end of the background chapter I will review the literature of US assessment tools and scope the aim of this thesis.

Description of the ultrasonography skill

US is a dynamic examination that differs from other stationary imaging modalities because the operator needs to combine both technical skills and image interpretation skills during the examination (43).

Technical skills in ultrasonography

Technical skills are needed to operate the US equipment and optimize the image according to the US examination (44). Further, technical skills are needed to move and manipulate the transducer guided by anatomy knowledge and the tactile information from the surface of the patient body (45). Appropriate amount of force needs to be applied to the transducer, e.g. to manipulate intestinal air in the abdomen in order to establish a good acoustic window. In this thesis I will use theories about motor control to explain the development of technical US skills. Motor skills are defined as activities that require voluntary movements to achieve a goal (46). In the context of US, the goal would be proper handling of the US equipment to perform a successful diagnostic US examination on patients. Motor control can be explained as memory-based motor programs that control the specific coordination of muscles needed to perform the procedure (46). However, the stationary description of motor control fails to explain how performance is adapted to the various clinical contexts where US is performed in real life. According to *schema theory* a motor program can be seen as a coordination concept for the movements of the particular procedure where different parameters (e.g. speed and amplitude of the movement) can be adjusted to the different situations (47). When the physician gains experience through clinical performance, the adjustment of these parameters to the variation of patients would be stored as a set of *schemas* in long term memory that can be used for controlling the performance in future situations. Gentile further developed a classification continuum of motor skills from closed skills to open skills based on four categories: *regulatory conditions* (stationary or in-motion) and *inter-trial variability* (absent or present) (46). *Regulatory conditions* describe changes in the environmental context that influence the movements needed to perform the procedure successfully. If the regulatory conditions are "in mo-

tion" it means that the environmental context changes during performance opposite to a "stationary" condition. The *inter-trial variability* describes if the similarities of the regulatory conditions are the same (absent) or changes from one trial to the next (present). Completely closed skills are stationary and can be performed the same way each time, while open skills require the performer to adjust to the changing environment during performance. This means that the motor program used to perform an US examination needs to be adjusted according to the patient to patient variation, i.e. inter-trial variability. In addition, US is a dynamic examination and the performance must be adjusted according to the position of the patient and movements of the body and organs, e.g. lung movements displacing the abdominal organs affecting the movement of the transducer in multiple planes, i.e. in-motion regulatory conditions. US can therefore be classified as a complete open motor skill where we can expect inter-case variability between the US cases. This is important in relation to assessment of skills because the use of a single case for performance-based assessment will not necessary predict successful performance of future clinical cases.

Image interpretation

Besides the technical skills required to perform the US examination, the physician also needs to correctly interpret the US images generated. Image interpretation can be divided into *visual perception* (visually search of the image) and *image analysis* (synthesis of image information into a conclusion about the diagnosis) (48). Spatial abilities are required to translate a 2-D sectional US image into a 3-D mental representation of the structure (45) and the visual perception can be categorized into two different search strategies: an *initial glance* and a *focal search* (48). The *initial glance* of the image gives the physician a global impression that is compared with information about normal anatomy and pathology stored in long-term memory of the physician. The abnormalities may stand out for the experienced physician who subsequently can make a quick decision about the diagnosis. The *focal search* is an interactive process where the image is systematically searched for features that attract attention and need further relevant information extracted from the image (48). Experts tend to diagnose images by the initial glance method within seconds, while novices rely on the focal search to establish a diagnosis (49). It is hypothesized that false-negative diagnostic errors can be divided in three categories: *search errors*, *recognition errors*, and *decision errors* (48). *Search errors* occur when the lesion is never fixed by the fovea of the eye. *Recognition errors* occur when the eye is shortly fixed on the lesion but below the threshold sufficient to recognize the pathology features of the image. *Decision errors* occur when the lesion is fixed by the eye but actively dismissed as a lesion. A decision of diagnosis is made if image features match sufficiently with the cognitive schema of abnormalities for a given disease. However, these concepts are mainly established on the exploration of the search strategies applied to x-ray images (48) and may not apply to the dynamic US examination where the physicians perform the US examination and interpret the image at the same time (50). This will increase the risk of *search errors* because a lesion may not be represented to the eye due to suboptimal scanning technique. We therefore need to combine both technical skills and image interpretation as one construct in the assessment of point-of-care US skills in order to ensure full competence of the physician.

Theory used in performance-based assessment

Assessment in health profession education can be defined as any systematic method of obtaining information from tests of learning and skill acquisition (51). According to Miller's taxonomy, competence can be divided into four levels with increased authenticity for assessment: 'knows', 'knows how', 'shows how' and 'does' (52).

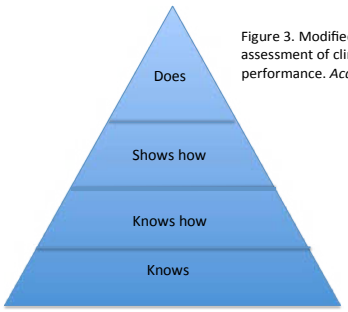

Figure 3. Modified from Miller, G. E. (1990). The assessment of clinical skills/competence/ performance. *Academic Medicine*, *65*(9), S63.

Knowledge is the foundation for skill development. Factual knowledge constitutes the 'knows' level in Millers pyramid and further understanding of concepts is needed to demonstrate a "knows how" level, see figure 3. Written and/or oral examinations are often useful for assessment of knowledge. The 'shows how' can be assessed under controlled conditions, e.g. with the Objective Structured Clinical Examinations (OSCE) while the 'does' level indicates the highest level of assessment with the physician functioning independently in clinical practice (51). However, it is important to differentiate between assessment of performance and learning. Learning is defined as a relatively permanent change in the ability to perform a skill (53) and performance during or immediately after training does not always predict sustainable learning (54). Skills learning should therefore be ex- plored by assessing performance after a retention period or by a transfer test–where the skills are performed in another setting (46). Further, skills assessment has to yield reproducible and accurate measurements before it can be clinically useful (55).

Physical quantities measured in health sciences (like temperature or a hemoglobin level) are often well-defined constructs measured in a reproducible fashion. The situation is different within behavioral science where the assessment is depending upon the definition of the construct (56): e.g. assessment of the patient's 'quality of life' or the physician's 'US skills' and the type of measurement instrument. In case the instrument is depending on a rater's judgement, the interpretation of construct being assessed may vary from one rater to another and thereby also the way a performance is being judged. That implies a threat to the validity of the assessment. Therefore, assessment of US skills needs thorough validation studies to ensure the best possible interpretation

of the assessment score. Different theoretical frameworks can be used to explore the validity evidence of performance assessment as discussed in the following section.

Reliability

The reliability of an assessment tool is the consistency and reproducibility of its measure between cases and raters over time (57). According to classic test theory, an assessment score can be decomposed in two parts: a true score and an error score (58). The true score would be the perfect measure of the competence and can only be obtained if there is no error in the measurement– which would never be the case in real life. The reliability coefficient is an estimation of the ratio between true score and error score in the assessment. Several factors contribute to error in the measurement and can decrease the reliability coefficient: tendency of a rater to mark differently compared to other raters (inter-rater variation); variation that occur from the same rater marking differently on two assessments of the same case over time (intra-rater variation); variations in performance owing to different assessment occasions (inter-case variation). Other unknown factors and the interactions between the factors can also contribute to error in the measurement. A critique of classic test theory is that it only provides a reliability coefficient from a single factor per analysis and cannot estimate the total error in the measurement (58). Generalizability Theory can instead be used to estimate the impact of the different factors contributing to error in the measurement of the same construct – allowing a more comprehensive assessment of reliability (51). Generalizability Theory defines 'true variance' as the differences in score between test objects that are stable across different cases and raters, while 'error variance' is the variation from all other sources (rater, cases, and interactions between them) (59). The variance compo- nents can be estimated in a G study to an overall Generalizability Coefficient (ranging from 0-1), expressing the percentage of the score attributable to true score. Further, the estimated variance components can be used in a Decision Study to estimate the number of raters and cases needed for reliable measurement (59).

Validity

Validity is traditionally divided into the trinitarian Cs of *content validity*, *criterion validity*, and *construct validity* (55). Content validity refers to the theoretical relationship between the content of the assessment tool and important aspects of the construct of interest. Often experts in the field are used to develop a test blueprint and cases representative of the construct. Criterion validity is defined by the correlation of an assessment tool against an accepted existing measure with well-known characteristics as a 'gold standard'. Construct validity refers to a collection of information that the assessment tool measures the construct it is supposed to measure. This can be difficult to determine and is typically explored by comparing assessment scores between groups with different levels of skills experience (60). The nomenclature of these distinct types of validity has been criticized to be arbitrary and instead Messick reconceptualized validity to a unitary concept where all sources of validity testing encompass construct validity (61). He defined five sources of validity evidence: *test content*, *response processes*, *internal structure*, *relations to other variables,* and *consequences*. The validation process involves a hypothesis testing with accumulation of different sources of evidence to support or reject the proposed test score interpretation in the particular setting. However, conclusions about validity are not dichotomous and different researchers looking at the same collection of validity evidence may arrive to different conclusions (62). Three of the validity sources according to Messick's unitary framework of validity: *test content*; *internal structure*; and *relations to other variables*, correspond to the traditional validity types: *content validity*; *reliability*; and *construct validity*, respectively. However, *response process* and *consequences* have not been described as validity evidence in the traditional framework. *Response process* is evidence regarding the control of error associated with the test administration (51). This could be done by examining the reasoning processes of learners to reduce possible response error and putting explicit and clear anchors on the rating scale or provide rater training to avoid misinterpretations among the raters (40). *Consequences* refer to the assessments scores' impact on and consequences for the individual and the society. A standard setting needs to be conducted with scientific argumentation for the pass-fail score of the assessment and an evaluation of the consequences of false positive and false negative outcomes. It is an important source of validity evidence (40) which unfortunately often is omitted in validity studies (63,64). Many different standard setting methods can be chosen (65) and I will therefore look in closer detail at the different methods in the next paragraph.

Standard setting methods

Standard setting is the process of defining the skill level required and establishing the corresponding assessment cut score defining the competence (65). Either normative or criterion-based approaches can be used for standard setting. With normative standards the pass/fail status depends on the performance of the tested cohort and this approach is therefore not optimal for competence-based education using the same standards for performance (66). Instead criterion based standard setting methods can be used to set an absolute assessment score that corresponds to the performance expectations of the particular skill by the passing physician (67). The criterion based standard setting can be further divided in either test-centered or examinee- centered methods (68). The Angoff, Ebel, and Nedelsky methods are examples of test- centered methods using the raters to decide pass-fail standards based on the examination content (65). In contrast, the examinee- centered methods focus on the performance of examinees to establish standard setting (69). A contrasting-groups approach is one way to define the standard for pass/fail with use of external criteria's to categorize the physicians into groups of competent vs non-competent. The test score that best discriminates between contrasting groups are then used as cut score (66). However, several statistical techniques can be used for establishing the optimal pass/fail score (70). One popular method is to plot

a “graphic smoothing” (70) where the pass/fail score is established by plotting the distributions of test scores for the competent and non-competent groups and using the intersect as pass/fail score (65,67,68,71-73), see figure 4A.

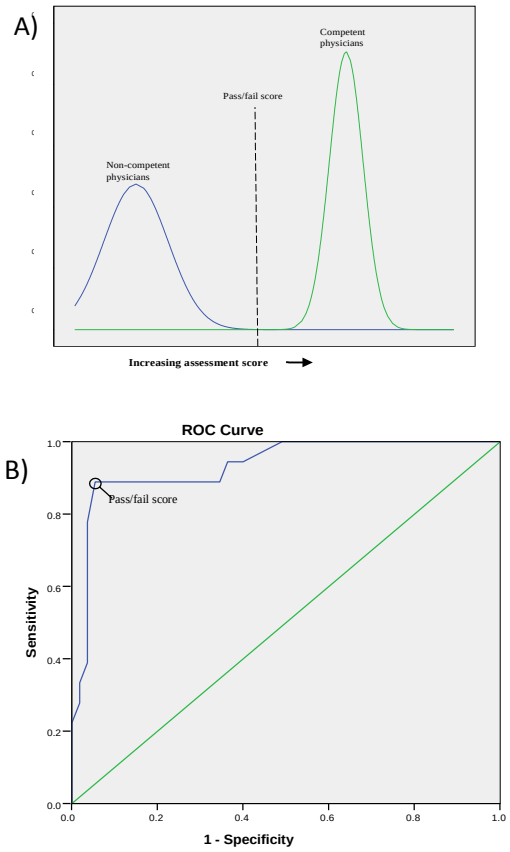

Figure 4. *Examples of two different techniques for deciding the pass/fail score between contrasting groups:*
A) Graphically decided. B) Based on a receiver operating characteristic (ROC) approach.

Another method–which is more unconventional for medical education research but commonly used in radiology to compare different diagnostic modalities–is the receiver operating characteristic (ROC) approach to establish the optimal cut score between two groups (74). The ROC plot provides a statistical method to assess the consequences of the cut score choice as a tradeoff of the true-positive rate (sensitivity) and false-positive rate (1– specificity) (75), see figure 4B. The prediction of the pass/fail rates related to the choice of cut score can be useful because standard setting depends on the purpose of the test as well as practical and economic considerations. However, we need to define which physicians have the necessary US competence before we can define a pass-fail score with the contrasting groups method. For this purpose we can use Dreyfus and Dreyfus' theory about skill acquisition to define when competency is achieved (76).

The Five-Stage Model of Skill Acquisition
The theoretical framework by Dreyfus and Dreyfus has been applied to the acquisition of clinical skills by physicians in training (77). It describes how the development from novice to expert level can be divided in five stages:
Novice: The novice physician has theoretical knowledge about the skill but little or no clinical experience. He has little ability to prioritize clinical information and the decisions are based on theory-based rules rather than adjusting skill performance to the environment in which it is performed.

Advanced beginner: After the physician gains some experience he starts to develop an understanding of the relevant information from the clinical situation and modify performance according to this.

Competent: The competent physician now sees the big picture and can operate in the context of a changing learning environment. He has developed a repertoire of illness scripts from clinical experiences to guide performance rather than rely on rule-based decisions. However, the competent physician has no experience with complex problems, which are handled in a rule-based way.

Proficient: The physician has enough experience to recognize and handle most common cases by intuition and is better than a competent physician to change plans according to the clinical situation. However, there is still room for improvement when complex cases are handled.

Expert: The expert physician uses intuition to solve the clinical problems because the pattern recognition is highly developed from prior experiences. However, at the same time the expert is also aware of when the unexpected may occur and performance therefore needs to be slowed down for a more careful and analytic approach. At this stage it is important for the expert physician to constantly improve her skills by using the extra mental resources from performance at automaticity level to invest them in more complex problem solving (77).

Different development stages from the Five-Stage Model of Skill Acquisition can therefore be used as baseline for the certification demands in concordance with the test consequences. According to Dreyfus and Dreyfus's model, the surgeon with US skills at a *competent* level would effectively diagnose common clinical cases, while complex cases instead need to be supervised by a senior consultant or referred to a radiologist. If a proficient level was chosen for certification we would expect the physician to handle more complex US cases, but still not at an expert level.

Assessment of US performance
Many of the existing US assessment tools found in the literature are developed to assess skills in US-guided invasive procedures and they primarily focus on the technical skills and not the image interpretation (78-86). No assessment tools were found specific for skills in head & neck US, while a few studies described assessment tools for abdominal US skills, especially specific for FAST competence (6,87-95). However, these tools are based on self-assessment questionnaires of competence (6,94,95) or either

assess only the technical skills (87-90) or US image interpretation skills (91-93) and not the combination. Instead Hofer et al. developed an assessment tool of both technical and interpretation skills of abdominal US (96). However, the OSCE test setup used was very comprehensive with a procedure specific checklist for each US examination of the abdominal organs and therefore requires a lot of rater training and resources. Further, only inter- nal validity evidence was explored for the assessment and the technical and interpretation skills were assessed on separate skill stations. This may not be optimal when the relationship between technical US skills and interpretation errors in US are taken into consideration and it can therefore be a threat to validity due to differences in the content of the assessment and the construct of interest. In the search for a solution to this problem we devel- oped the Objective Structured Assessment of Ultrasound Skills (OSAUS) scale together with Tolsgaard et al. (97). The content of the generic scale obtained international consensus from experts representing multiple specialties using US in diagnostic of pa- tients. The OSAUS scale covered both technical skills and image interpretation through seven key elements of the US examination assessed on a five-point Likert scale. It should therefore cover the complete US skill and by feasible to implement for in-training assessment of surgeon-performed US skills. However, only validi- ty evidence regarding test content was established in this study (97). According to Messick's unitary framework of validity (98) the other sources of validity evidence regarding response processes, internal structure, relations to other variables, and consequences still need to be established.

In conclusion, US is an user-dependent diagnostic modality with requirement of both technical and image interpretation skills of the operator. Development of portable US equipment has facilitated the use of surgeon-performed US and increased the need for competence-based training to ensure safe integration of US into clinical practice. An assessment tool for both technical and image interpretation skills of the US operator is needed and this thesis aimed to explore the validity evidence of the OSAUS scale to assess skills in abdominal and head & neck US.

OBJECTIVES OF THE THESIS

This thesis aimed to collect validity evidence regarding response processes, internal structure, relations to other variables, and consequences of skills assessment in surgeon-performed abdominal and head & neck US with the use of the OSAUS scale.

Research questions

Five research questions were generated for this PhD thesis. The aim of *research project I* was to collect evidence of validity related to internal structure and relationship to other variables of the OSAUS scale for assessment abdominal point-of-care US skills. The research questions were:

1) "What is the reliability and how many cases and raters are needed for reliable judgment of physicians' point-of-care US competence using the OSAUS scale?" (1)
2) "What is the validity of the OSAUS scale in terms of its ability to discriminate between increasing levels of US competence and association between OSAUS scores and diagnostic accuracy? "(1)

The aim of *research project II* was to investigate the transfer of skills learned from an abdominal point-of-care US course to performance on patients representing pathological conditions. The research question was:

3) In a group of clinicians, what is the effect of participating in a four-hour course in abdominal point-of-care US as measured by OSAUS score on patients representing abdominal diseases compared with having no training?

The aim of *research project III* was to explore the internal structure, relations to other variables, and consequences of the OSAUS scale for assessment of head & neck US skills in an office-based setting. The research question were:

4) Can reliable assessment of head & neck US skills with the OSAUS scale be obtained with the use of raters from different specialties?
5) What is the association between the OSAUS scores and the diagnostic accuracy of patients with or without head & neck lesions, and what are the consequences of establishing an OSAUS pass/fail score?

DESCRIPTION OF THE RESEARCH PROJECT

This section presents a brief summary of trial design and statistical methods used in the three research projects constituting this thesis.

Research project I

Trial design: The experimental study was conducted at the Clinical Skills Laboratory at Copenhagen Academy for Medical Education and Simulation, Rigshospitalet, Denmark (formerly known as Centre for Clinical Education). It was designed to establish validity evidence regarding the internal structure and relations to other variables of the use of the OSAUS scale to assess skills in abdominal point-of-care US. Twenty-four physicians were included in the study and were divided in three groups based on their experience with US: novices, intermediates, and experts. All physicians had to perform a focused US examination of three patients with abdominal pathology and one simulated patient with normal anatomy. They were allowed five minutes to complete each of the ultrasound examinations and make a short dictation of their findings to the medical record. The US performances were video recorded and merged with the ultrasonography screen recordings to form an anonymized clip for each case. All the video clips were uploaded to the Integrable web-based Solution for Easy Assessment (ISEA) web-based program for assessment of video recorded performance (99) from where two consultant radiologists assessed the US performance with use of the OSAUS scale.

Statistical methods: Data regarding the internal structure of OSAUS were explored with Generalizability theory estimating the relevant variables that influenced the reliability and a D-study determining the number of US cases and raters needed to make a reliable assessment (100). Kruskal-Wallis and Mann-Whitney tests were used to explore the OSAUS score's relations to US experi- ence levels and the correlation to diagnostic accuracy was ex- plored with Spearman ρ.

Research project II

Trial design: The second study was also conducted at the Clinical Skills Laboratory at Copenhagen Academy for Medical Education and Simulation, Rigshospitalet, Denmark. Physicians who signed up for an abdominal point-of-care US course were invited to participate in this study. Thirty-one physicians were enrolled in the study and randomized for assessment of their US skills before (control group) or after (intervention group) completing the US course. The test setup was similar to the one used in *research paper I* and US performance was also assessed by the same two radiologists using the OSAUS scale through the ISEA program (99). Some of the physicians from the control group also provided data to the novice group in *research paper I*.

Statistical methods: The internal structure of the OSAUS scale was explored with inter-rater reliability calculated by Cronbach's alpha. The relation to other variables was explored with an inde- pendent samples t test comparing OSAUS scores between the intervention and control group. The effect size of the learning from the US course was estimated using Cohen's d – with 0.2 representing a small ES, 0.5 a medium ES, and 0.8 a large effect size (101). Binary logistic regression was used to compare the difference in the diagnostic accuracy between the control and intervention group and generalized estimating equation was used to adjust for clustering of cases within each physician.

Research project III

Trial design: An experimental study was conducted at the out- patient clinic at the Department of Otorhinolaryngology, Head and Neck Surgery & Audiology, Rigshospitalet, Denmark.

Six US experienced otolaryngologist–head & neck surgeons (ORL-HNS) and 11 interns were included in the study. The participants had to perform neck US examinations at eight different test sta- tions–including six patients with verified neck lesions and two simulated patients with no evidence of disease. At each station the physicians were given a description of the patient's symp- toms, before they were allowed four minutes to complete a fo- cused US examination and additional four minutes to write their US report. The US performances were recorded in five of the eight patient stations, and merged with the ultrasonography screen recordings to form an anonymized clip for each case. A consultant in diagnostic radiology and a consultant in ORL-HNS assessed all the video clips through the ISEA program with use of the OSAUS scale. The diagnostic accuracy was calculated as the percentage of correct diagnoses based on an evaluation of the US reports by a blinded ORL-HNS.

Statistical methods: The internal structure of the OSAUS scale was explored with the inter-rater and inter-case reliability of the assessment scores calculated with intra-class correlation coeffi- cient (ICC), average measures, with absolute agreement defini- tion. The relation to other variables was explored with a Spear- man ρ correlation coefficient between the OSAUS score and the diagnostic accuracy. The participants who met the criteria as novices or competent in head & neck US were included in a standard setting to establish an OSAUS pass/fail score. The con- sequences of the test was explored with a Receiver Operator Characteristic curve used to establish the optimal OSAUS pass/fail score and the area under the curve (AUC) was calculated and used to interpret the discrimination ability of the OSAUS score.

SUMMARY OF THE RESULTS AND RELATION TO OTHER RE- SEARCH

In this section the results from the three research papers will be summarized with the use of the sources of validity evidence from Messick's unitary framework explored in this PhD thesis: *response processes*; *internal structure*; *relations to other variables*; and *consequences*.

Response Process

The US performances of the participants in the studies for this thesis were all video-recorded and assessed by the raters using the web-based ISEA program (99). The program was designed to improve the response process by a standardized assessment form and automatic data generation to decrease the risk of error com- pared to the alternative where data had to be manually typed in. The response anchors on the OSAUS scale were included to avoid ambiguity and to ensure similar performance assessment through the use of the rating scale by different raters. We also performed structured rater training prior to data collection in all the studies of this Ph.D. thesis to further improve the validity evidence re- garding the response process. A similar rater training was per- formed in the study by Tolsgaard et al. (71) while the assessment tool regarding transthoracic echocardiography (102) and for ultrasound-guided endoscopic needle aspiration (81,83) did not describe how the rater training was performed. The amount of rater training will also influence the results from the internal structure and therefore make the comparison of the results be- tween the studies more difficult. The rater training in our studies consisted of a short introduction to the OSAUS rating scale and video examples of US performance used for discussion for pro- posed ratings to reach consensus. This was completed within an hour in all the studies in this PhD thesis, why we believe it is realistic to use the scale in a clinical setting as well. However, the validity evidence regarding the response process will of course change if the OSAUS scale is used for direct observation (e.g. for workplace-based assessment purposes) compared to the use of video assessments and blinded raters in our studies (81).

Internal structur

Generalizability theory was used in *Research paper I* to explore the internal structure of the OSAUS scale's ability to assess

physi- cians' abdominal point-of-care US skills. The generalizability coef- ficient was high (0.81) in a test setup using two raters and four different abdominal US cases, which is considered sufficient for high-stake assessments. If absolute agreement definition was used (e.g. to set an OSAUS cut score), five cases and two raters were needed to ensure sufficient reliability. The raw variance components from the generalizability analysis were also de- scribed in the *Research paper I* to ensure an appropriately sample and evaluation of all relevant factors for the internal structure (103).

Source of variance (V)	Description	Estimated variance component (VC)	Relative contribution	Interpretation of results
Physicians, Vp	Systematic variation among physicians	9.91	43.7%	Most of the measured variance derives from different competence between the physicians
Rater, Vr	Systematic variability among raters	1.30	5.75%	The raters had an overall equal level of stringency
Case, Vc	Systematic variability among cases	0.201	0.887%	The cases were almost equally difficult
Interaction between physician and rater, Vpr	Consistent trend for an rater to assess a particular physician differently	0.294	1.30%	There was no bias between rater and physician due to effectively blinding
Interaction between physician and case, Vpc	Consistent trend for an physician to perform a particular case differently	5.96	23.6%	Some variance derives from the physicians who score different according to the case
Interaction between case and rater, Vcr	Consistent trend for a rater to assess differently on a particular case	-0.0236	0.1%	The raters do not vary in their perceptions of the challenge of an ultrasound case.
Interaction between physician, case and rater, Vpcr	All remaining variability	4.99	22.0%	Expected unexplained error

Table 1. *Results from the G-study indicating the contribution of each source of variance to the OSAUS score (adjusted from Table 2 in Research paper I).*

The score variance that could be described as 'true variance' of competence, 43.7%, was lower compared to other Generalizabil- ity studies exploring assessment with generic scales of ultra- sound-guided transbronchial needle aspiration, 52.1% (83), and transthoracic echocardiography skills, 67% (83,102). The main reason for lower true variance in *research paper I* is the increased 'error variance' from the interaction between the physician and case (*Vpc*) on 23.6%[1] compared to 11.1% for assessment of endobronchial US skills (83) and 7.1% in echocardiography skills (102). However, these findings are not surprising as the assessment of abdominal US is a broad construct including many different organs. It is possible that a surgeon may have more experience with conducting FAST examinations com- pared to other abdominal US examinations, which will increase the variance due to US cases compared to a more unitary con- struct as an US guided interventional procedure (83) or echocar- diography (102). In contrast a D-study about an assessment scale used in the evaluation of ultrasound-guided transesophageal fine- needle aspiration skills found similar results compared to *research paper I* (103).

[1] Unfortunately an error was slipped in research paper I's Table 2 where the variance components from the interaction between physician and the case (*Vpc*) was interchanged with the interaction between rater and the case (*Vrc*). Instead the table should have shown that the *Vpc* contributed with 23.6% of the measure vari- ance while the *Vrc* did not contributed with any variance (see Table 1).

However, the study did not report the raw variance components from the skills assessment, making it difficult to directly compare with the results from *research paper I*. Validity evidence regarding the *internal structure* of the OSAUS scale to assess abdominal point-of-care US skills was also supported by good reliability from classic test theory analyses of the OSAUS scores in *research paper II*. Further, results from validity studies exploring the ability of the OSAUS scale to assess obstetrics and gynecology (71) and head & neck US skills (*research paper III*) also supported the validity of the internal structure.

Relation to other variables

The results from *research paper I* have shown that the OSAUS scale relation's to other variables was supported by both a signifi- cant difference in scores between physicians with different levels of point-of-care US experience, and by a strong correlation be- tween the OSAUS score and the diagnostic accuracy. We there- fore believe the OSAUS scale can be used to assess progress in abdominal point-of-care US skills during training. The ability of the OSAUS scale to discriminate between different experience levels with abdominal point-of-care US in *research paper I* was also established for US skills in both obstetrics and gynecology
(71) and head & neck surgery (*research paper III*). Further, the OSAUS scale's relation to other variables was supported by a good correlation to diagnostic accuracy in abdominal US (Spear- man ρ of 0.76 in *research paper I*) and head & neck US (Spearman ρ of 0.85 in research paper III). A recent study introduced a Quali- ty of Ultrasound Imaging and Competence (QUICk) assessment tool developed to assess US operator competence in FAST (90). The assessment tool consists of both a task-specific checklist and a global rating scale assessing US technical skills. Relation to other variables was also explored for the QUICk scale (90), but only in the form of discrimination of two extreme groups (US novices and US experts). Further, the inter-rater reliability was lower for the QUICk global rating scale measured as weighted kappa on 0.61, compared to an intraclass correlation coefficient on 0.86 (con- sistency definition) for the OSAUS scale in *research paper II*. De- spite different statistics used to explore the QUICk scale and the OSAUS scale with weighted kappa and intra-class correlation coefficient, respectively, the results should still be comparable (55). The content validation of the QUICk scale was not described in details and the assessment tool only included the technical aspects of the US examination and not the image interpretation, documentation, and medical decision making (104). This may explain the low inter-rater reliability of the QUICk scale compared to OSAUS. Further, the study only used one volunteer without pathology as test case, which is problematic regarding the con- struct of measurement as discussed in the background chapter.

Test consequences

The *research paper III* conducted a standard setting study that established a pass/fail score used for competency-based assess-

ment of head & neck US skills. The assessment format demonstrated good discrimination of US competence to ensure a baseline of diagnostic accuracy among physicians. The used ROC curve approach (105) in *research paper III* is a bit different than comparable validity studies performing standard setting for performance in endoscopic (106) and gynecological US (71). However, the OSAUS pass/fail scores of 3.0 and 2.5 for transabdominal and transvaginal US, respectively (71) is comparable to the OSAUS pass/fail score of 2.8 for head & neck US competence in *research paper III.* The choice of standard setting method may therefore not be so critical to the decision of exact pass/fail score. Instead, the ROC curve approach can be used to choose different pass/fail scores according to test purpose and consequences of the test (e.g. formative assessment during training versus summative assessment with certification for independent practice). Ziesmann et al also used ROC curve analyses to establish standard setting for operator competence in FAST assessed with the QUICk scale (90) and hand motion analysis (89). The area under the curve (AUC) in *research paper III* was comparable with the AUC results from the use of the QUICk scale (90) and better than using hand motion analysis (89) to measure US competence. However, the comparison of the AUC between studies can be arbitrary, because it also depends on the US skills level of the groups compared.

Summary of validity evidence

This thesis contributes with strong validity evidence of the OSAUS scale for assessment of US skills of abdominal and head & neck diseases. The summary of validity evidence can be organized with Messick's five different sources to support for the interpretation of the OSAUS assessment:

1. *Test content* was established in a prior Delphi study with international multi-specialty expert consensus about content of the OSAUS scale (97) and was not explored further in this thesis.
2. *Response process* of the assessment was ensured by rater training provided in the studies, an uniform and standardized testing procedure, and the development of the ISEA web-based assessment solution to automatize the processes of data entry (99).
3. *Internal structure* was supported by a reliability coefficient sufficient for high stakes examinations. Reliability was both explored with classic test theory and with Generalizability theory for a more comprehensive assessment of reliability and to estimate how many raters and test cases would result in similar reliability in future studies.
4. Relations to other variables were demonstrated by the OSAUS scale's ability to differentiate between experience-levels in US and measure effect of training. Further validity evidence was supported by a high correlation between the OSAUS score and the diagnostic accu- racy of both abdominal and head & neck US.
5. *Consequences* were explored by a contrasting group standard setting of the OSAUS pass/fail score to define competence in head & neck US. A receiver operator curve analyses demonstrated a good OSAUS score dis- crimination of head & neck US competence with ac- ceptable pass/fail test consequences.

Overall, this thesis collected validity evidence regarding response processes, internal structure, relations to other variables, and consequences to support the OSAUS scale for skills assessment in surgeon-performed abdominal and head & neck US.

Transfer from simulation training

The results from *research paper II* have shown that physicians could successfully transfer learning from an ultrasonography course combining didactic with hands on training on healthy human models to improved point-of-care US performance and diagnostic accuracy on patients with abdominal pathologies. However, the results indicate a need for more training before competence in abdominal point of care US is obtained. Other studies found good learning outcomes from US courses, but these studies only measured the technical abilities on healthy human models (107) or image interpretation by assessing US video clips (108-111). Further, other studies exploring the effect of formal US education lack assessment tools supported by validity evidence (6,112-114). *Research paper II* is therefore the first study to estab- lish the effect of formal abdominal US training measured by the diagnostic performance on patients with an assessment tool supported by strong validity evidence. In our study, the physicians who received 'hands on' training improved substantially on all OSAUS items except regarding image optimization. Low OSAUS scores on image optimization were also found by US experienced abdominal surgeons in *research paper I* and ORL–H&N surgeons in *research paper III*, especially when the score is compared with consultants in fetal medicine performing transabdominal fetal US (71). One explanation of our findings could be that the physicians in the study were unfamiliar with the US equipment available in the experimental studies and therefore did not optimize the US images.

However, it could also be due to lack of knowledge of this part of the US examination. When we compare the mean OSAUS score of the physicians who received 'hands on' training (27%) from *re- search paper II* with the expert group (71%) from *research paper I,* it is indicated that much more training is needed than a single formal course. The formal hands on US courses may therefore need to be longer to prepare physicians for clinical point of care US practice as suggested by other studies (108,115). Further research needs to explore how to ensure progress in US skills after formal US training is completed.

LIMITATIONS

Generalizability of findings

The experimental setups used in the studies of this PhD thesis ensured controlled test conditions to assess US performance with a minimum of confounders influencing the study outcome (60). However, these studies do not directly assess how the physicians perform point-of-care US in their daily clinical work, which is considered the top of Millers pyramid (116). We will therefore discuss some issues regarding the generalizability of our main results in this paragraph.

OSAUS ratings

Because the OSAUS elements regarding "indication for the examination" and "medical decision making" were not relevant to assess in the experimental studies of this thesis, only five of the seven OSAUS elements established from the Delphi consensus study (97) were used for assessment of US competence. We found it awkward to assess the "indication for the examination" because the physicians already were assigned to perform the US examination as a part of our test setup. The "medical decision making" in point-of-care US is depending on the history and clinical examination of the patient and therefore we did not include this as a part of the assessment in our test setups. Further, we may expect a difference in the medical decision making by radiologists compared to surgeons, which can compromise the reliability of the measurement of this OSAUS element. According to the OSAUS content validity study (97) these two elements should only be assessed if applicable to the concrete setting which we did not find it to be according to the aim of this thesis. Because the reliability will "artificially" improve simultaneously with increased number of scale items (55), we believe it is a strength to the OSAUS scale that we demonstrated good reliability regarding the use of only five items in this thesis.

Diagnostic accuracy

The restricted time in our test setup to complete the US examination may not be representative to the clinical setting where additional time can be used for more challenging US examinations. However, the time limit ensured equal test conditions for all the participating physicians and a focused point-of-care US should be completed within the allowed time in office-based setting. Further, the US reports in our studies were used to classify the diagnoses as correct or false in order to calculate the diagnostic accuracy. Sometimes the US examination can be inconclusive due to US artifacts, which is important to recognize by the US operator and should not be confused with an misinterpretation of the US image (115). In our studies the physicians were "forced" to make diagnosis of the patients they examined, while they in their real clinical work may have ordered another imaging or consult a colleague instead. The diagnostic accuracy found in our experimental study may therefore have been higher in a clinical setting for the surgeons. However, when our studies are compared to clinical studies exploring diagnostic accuracy of surgeon-performed US, they often used a single dedicated and intensively trained surgeon (117-120). These studies will therefore inflate the diagnostic accuracy compared to a real clinical setting where surgeons with varied US training and competence would conduct the US examinations. A study among trauma surgeons found sizable variations in the US diagnostic accuracy but attributed their findings to patient variation (121). The strength of our studies is the measure of individual differences in diagnostic accuracy in a controlled setting.

Study design

As mentioned in the background of this PhD thesis, reliability is established as an interaction between the variance of the subjects and the error variance of measurements. In *research paper I and III* we artificially increased the magnitude of variance of the subjects by using novices, intermediate and experts in US as research objects, which therefore will increase the reliability coefficient (62). If the assessment is applied to a more homogeneous population of physicians, e.g. surgical residents during training, we would expect the reliability of the measurement to decrease.
Further, by comparing extreme groups we do not explore the 'responsiveness', i.e. the ability of the scale to measure a meaningful clinical change as progress in US skills among physicians (62) or differentiate between US experienced physicians with or without competence for independent US practice. However, in *research paper II* we demonstrated the ability to measure the effect of an educational intervention, which also can be seen as a proof of responsiveness of the OSAUS scale to measure progress in US competence. Further, validity evidence is context specific and our results obtained from an experimental study with use of video recorded US performance may not directly generalize to a clinical setting with direct observation by a faculty member (81).

Training vs no training

In *research paper II* we conducted a randomized controlled trial to explore the transfer of skills from an abdominal point-of-care training course to diagnostic performance of patients. Although the choice of a randomized controlled clinical trial is recognized as the highest level of clinical evidence (122), our comparison between an intervention group receiving US course training to a control group receiving no training is problematic in educational research (123). This only tells us that learning can occur when learners spend time on training, and does little to inform us how to improve educational practice (124). However, transfer of skills to improve patient care cannot be taken for granted and many educational interventions fail to change the clinical performance of physicians (125,126). Further, many surgeons use point-of-care US without any formal training (127), why we find it important to establish evidence for the effect of US courses.

IMPLICATIONS

The results from this thesis can be used to guide the development of a competence-based education of surgeons in US. The OSAUS scale can be used to assess competence instead of the one-size-fits-all approach with requirements of specific number of completed US examinations. The progress in US skills can thereby be followed by in-training OSAUS assessments during supervision of the performance (128) or by uploading US clips for assessment online (99,129). If the OSAUS score is based on a single case assessed by a single rater the reliability of the assessment would not be impressive according to our D-study in *research paper I*, but still it is sufficient for the formative assessment purpose to evaluate progress in US skills. The latest 2015 updates of the otolaryngology-head & neck surgery residency program in Denmark recommended the use of the OSAUS scale as assessment tool for head & neck US skills (38). However, no specifications regarding number of assessment cases or a minimum OSAUS score needed for certification were clarified. *Research paper III* established a pass/fail score that can be integrated as part of certification process to ensure quality of the US scans provided by trainees in head & neck surgery. To ensure the reliability (G-coefficient > 0.8) of high-stakes decisions like certification for

independent practice, five cases and two raters are needed. However, the OSAUS pass/fail score should not be used as the only measure of when a trainee is ready for independent practice, but rather be used alongside other markers of competence as well. Although much more training is needed after course completion, this thesis found good effect of formal US training courses on US performance and it should therefore constitute the basis of the US training before clinical practice. Especially US image optimization in general needs to be improved when surgeons use US in gynecology (71), abdominal (research paper I and II) and head & neck surgery (research paper III). Poor image optimization may be due to lack of theoretical knowledge and we therefore recommend a competence-based education to include US knowledge assessment to ensure the fundament to clinical skills(130). Thereafter the OSAUS scale can be used to measure progress in US performance while an OSAUS pass/fail score can be used as final certification.

CONCLUSIONS AND PERSPECTIVES FOR FURTHER RESEARCH

This thesis demonstrated strong evidence supporting the interpretation of the OSAUS scores as measure of physicians' abdominal and head & neck US competence. The thesis established recommendations for the optimal administration of assessment of US skills with the use of the OSAUS scale and defined pass/fail standards of US performance. We found that physicians successfully could transfer learning from an ultrasonography course to improved point-of-care US performance and diagnostic accuracy on patients. However, it was also demonstrated that much more training is needed after formal course training and surgeons need to improve the image optimization of their US examination. The results from this thesis can both be integrated in residency training and certification and is therefore an important step towards competency-based education in surgeon-performed US.

The amount of training needed to gain competence in surgeon-performed US will vary and further research needs to explore how courses and clinical training should be organized to optimize the US training. How to use new technologies like US simulators (131,132) and e-learning (133,134) to improve the surgical US education is also a scope for future research. Further, our studies established the OSAUS assessment to ensure the diagnostic accuracy of US under controlled conditions and translational studies should explore if it can be integrated into assessment of the clinical performance (60). More research therefore needs to explore how the use of point-of-care US will affect the medical decision making by surgeons, and finally how it will affect patient outcome and cost-effectiveness of the health care system (135).

LIST OF ABBREVIATIONS

FAST	Focused Assessment of Sonography for Trauma
ORL-HNS	Otolaryngologist–head & neck surgeon
OSAUS	Objective Structured Assessment of Ultrasound Skills
OSCE	Objective Structured Clinical Examinations
US	Ultrasonography / Ultrasound
ISEA	Integrable web-based Solution for Easy Assessment of video-recorded performances
QUICk	Quality of Ultrasound Imaging and Competence

SUMMARY

Surgeons are increasingly using ultrasonography (US) in their clinical management of patients. However, US is a very user-dependent imaging modality and proper skills of the US operator are needed to ensure quality in patient care. This thesis explores the validity evidence for assessment of competence in abdominal and head & neck ultrasonography using the Objective Structured Assessment of Ultrasound Skills (OSAUS) scale. With the use of Messick's unitary framework of validity, five sources of validity evidence were explored: *test content*, *response processes*, *internal structure*, *relations to other variables,* and *consequences*. *Research paper I* examined validity evidence for the use of the OSAUS scale to assess physicians' abdominal point-of-care US competence in an experimental setting using patient cases with and without pathological conditions. The results provided validity evidence of the internal structure of the OSAUS scale and a decision study predicted that four cases and two raters or five cases and one rater could ensure sufficient reliability in future test setups. The relation to other variables was supported by a significant difference in scores between US experience levels, and by a strong correlation between the OSAUS score and diagnostic accuracy. *Research paper II* explored the transfer of learning from formal point-of-care US training to performance on patients in a randomized controlled study. The results supported validity evidence regarding OSAUS scores' relation to other variables by demonstrating a significant discrimination in the progress of training–a more refined validity evidence than the relation to difference experience levels. The results showed that physicians could transfer the skills learned on an ultrasonography course to improved US performance and diagnostic accuracy on patients. However, the results also indicated that following an initial course, additional training is needed for physicians to achieve competence in US. *Research paper III* evaluated validity evidence supporting an OSAUS score used to establish pass/fail standards for head & neck US skills. Good reliability between raters from different specialties to assess head & neck competence further supported the internal structure of OSAUS scale. A strong correlation to the diagnostic accuracy supported the relation to other variables and the consequences of the assessment were explored by a receiver operator characteristic curve for different pass/fail standards of head & neck US skills.

In summary this PhD thesis established sources of validity evidence supporting the interpretation of the OSAUS scale to evaluate surgeon-performed US skills of the abdominal and head & neck diseases. We therefore recommend the OSAUS scale for formative in-training assessment and high-stakes summative decisions as certification for independent practice in surgeon-performed US. Further, we find formal "hands on" courses an essential part of initial US training with good transfer of learning to improved diagnostic accuracy. This thesis can therefore be used to support the move towards competency-based training in abdominal and head & neck US.

REFERENCES

4. Newman PG, Rozycki GS. The history of ultrasound. Surgical Clinics of NA. 1998 Apr;78(2):179–95.

5. Merritt CR. Ultrasound safety: what are the issues? Radiology. 1989 Nov;173(2):304–6.

6. Webb EM, Cotton JB, Kane K, Straus CM, Topp KS, Naeger DM. Teaching point of care ultrasound skills in medical school: keeping radiology in the driver's seat. Academic Radiology. 2014 Jul;21(7):893–901.

7. Stenman C, Thorelius L, Knutsson A, Smedby Ö. Radiographer-acquired and radiologist-reviewed ultrasound examination--agreement with radiologist's bedside evaluation. Acta Radiol. SAGE Publications; 2011 Feb 1;52(1):70–4.

8. Woo MY, Frank JR, Lee AC. Point-of-care ultrasonography adoption in Canada: using diffusion theory and the Evaluation Tool for Ultrasound skills Development and Education (ETUDE). CJEM. Cambridge University Press; 2014 Sep 1;16(05):345–51.

9. Moore CL, Copel JA. Point-of-care ultrasonography. N Engl J Med. 2011 Feb 24;364(8):749–57.

10. Morris AE. Point-of-care ultrasound: seeing the future. Curr Probl Diagn Radiol. 2015 Jan;44(1):3–7.

11. Lewiss RE, Tayal VS, Hoffmann B, Kendall J, Liteplo AS, Moak JH, et al. The core content of clinical ultrasonography fellowship training. Acad Emerg Med. 2014 Apr;21(4):456–61.

12. Mazzaglia PJ, Milas M. Ultrasound training among endocrine oncology surgeons: what is best practice? http://dxdoiorgepfjernadgangkbdk/102217/ije158. Future Medicine Ltd London, UK; 2015 May 8;2(2):111–7.

13. Rozycki GS, Ochsner MG, Jaffin JH, Champion HR. Prospective evaluation of surgeons' use of ultrasound in the evaluation of trauma patients. J Trauma. 1993 Apr;34(4):516–26–discussion526–7.

14. Inaba K, Chouliaras K, Zakaluzny S, Swadron S, Mailhot T, Seif D, et al. FAST Ultrasound Examination as a Predictor of Outcomes After Resuscitative Thoracotomy: A Prospective Evaluation. Annals of Surgery. 2015 Sep;262(3):512–8.

15. Lindelius A, Törngren S, Pettersson H, Adami J. Role of surgeon-performed ultrasound on further management of patients with acute abdominal pain: a randomised controlled clinical trial. Emergency Medicine Journal. 2009 Aug;26(8):561–6.

16. Léger P, Fleet R, Maltais-Giguère J, Plant J, Piette É, Légaré F, et al. A majority of rural emergency departments in the province of Quebec use point-of-care ultrasound: a cross-sectional survey. BMC Emerg Med. BioMed Central; 2015;15(1):36.

17. Czerwonka L, Freeman J, McIver B, Randolph GW, Shah JP, Shaha AR, et al. Summary of proceedings of the second World Congress on Thyroid Cancer. Head Neck. 2014 Jul;36(7):917–20.

18. St J Blythe JN, Pearce OJ, Tilley EA, Brennan PA. Contemporary use of imaging modalities in neck mass evaluation. Atlas Oral Maxillofac Surg Clin North Am. 2015 Mar;23(1):1–14.

19. Esen G. Ultrasound of superficial lymph nodes. Eur J Radiol. 2006 Jun;58(3):345–59.

20. Katz P, Hartl DM, Guerre A. Clinical ultrasound of the salivary glands. Otolaryngol Clin North Am. 2009 Dec;42(6):973–1000–TableofContents.

21. Bumpous JM, Randolph GW. The expanding utility of office-based ultrasound for the head and neck surgeon. Otolaryngol Clin North Am. 2010 Dec;43(6):1203–8–vi.

22. Norling R, Buron BMD, Therkildsen MH, Henriksen BM, Buchwald von C, Nielsen MB. Staging of cervical lymph nodes in oral squamous cell carcinoma: adding ultrasound in clinically lymph node negative patients may improve diagnostic work-up. PLoS ONE. Public Library of Science; 2014;9(3):e90360.

23. Yeh MW, Bauer AJ, Bernet VA, Ferris RL, Loevner LA, Mandel SJ, et al. American Thyroid Association statement on preoperative imaging for thyroid cancer surgery. Vol. 25, Thyroid : official journal of the American Thyroid Association. 2015. pp. 3–14.

24. Welkoborsky H-J. Ultrasound usage in the head and neck surgeon's office. Curr Opin Otolaryngol Head Neck Surg. 2009 Apr;17(2):116–21.

25. Goldenberg E, Gilbert BR. Office ultrasound for the urologist. Curr Urol Rep. Current Science Inc; 2012 Dec;13(6):460–6.

26. Toprak H, Kiliç E, Serter A, Kocakoç E, Ozgocmen S. Ultrasound and Doppler US in Evaluation of Superficial Soft-tissue Lesions. J Clin Imaging Sci. 2014;4:12.

27. Nagarkatti SS, Mekel M, Sofferman RA, Parangi S. Overcoming obstacles to setting up office-based ultrasound for evaluation of thyroid and parathyroid diseases. Laryngoscope. 2011 Mar;121(3):548–54.

28. Carroll WW, Walvekar RR, Gillespie MB. Transfacial ultrasound-guided gland-preserving resection of parotid sialoliths. Otolaryngol Head Neck Surg. 2013 Feb;148(2):229–34.

29. Wanzel KR, Ward M, Reznick RK. Teaching the surgical craft: From selection to certification. Curr Probl Surg. 2002 Jun;39(6):573–659.

30. Touchie C, Humphrey-Murto S, Varpio L. Teaching and assessing procedural skills: a qualitative study. BMC Med Educ. BioMed Central Ltd; 2013;13(1):69.

31. The American Institute of Ultrasound in Medicine.

AIUM Practice Guideline for the Performance of Ultrasound Examinations of the Head and Neck. J Ultrasound Med. 2014 Feb;33(2):366–82.

32. Committee EAPS. European Federation of Societies for Ultrasound in Medicine and Biology. Minimum training recommendations for the practice of medical ultrasound. Ultraschall Med. 2006 Feb;27(1):79–105.

33. European Federation of Societies for Ultrasound in Medicine and Biology, Biology. Mimimum training requirements for the practice of medical ultrasound in Europe [Internet]. [cited 2016 Feb 7]. Available from: http://efsumb.org/guidelines/2009-04-14apx5.pdf

34. The American Institute of Ultrasound in Medicine. Training Guidelines for Physicians Who Evaluate and Interpret Diagnostic Abdominal/General Ultrasound Examinations [Internet]. [cited 2016 Feb 7]. Available from: http://www.aium.org/resources/viewStatement.aspx?id=47

35. American Institute of Ultrasound in Medicine. Training Guidelines for Physicians Who Evaluate and Interpret Diagnostic Ultrasound Examinations of the Head and Neck [Internet]. aium.org. [cited 2016 Feb 7]. Available from: http://www.aium.org/resources/viewStatement.aspx?id=55

36. Deutsche Gesellschaft für Ultraschall in der Medizin. Antrag auf Anerkennung der DEGUM-Stufe II Sektion Kopf-Hals (HNO, MKG, Kopf-Halsradiologie) [Internet]. [cited 2016 Feb 7]. Available from: http://www.degum.de/fileadmin/dokumente/sektionen/kopf-hals/Antraege/Antr%C3%A4ge_11_2015/KPF_Antrag_Stufe_II_ 2015.pdf

37. The Danish Health Authority. Målbeskrivelse for speciallægeuddannelsen i Kirurgi [Internet]. [cited 2016 Apr 25]. Available from: https://sundhedsstyrelsen.dk/da/uddannelse/speciallaeger/maalbeskrivelser/~/media/E887147BCEDF442AA8BF068D3F54CCB6.ashx

38. The Danish Health Authority. Målbeskrivelse for Speciallægeuddannelsen i Oto-Rhino-Laryngologi [Internet]. [cited 2016 Apr 25]. Available from: https://sundhedsstyrelsen.dk/da/uddannelse/speciallaeger/maalbeskrivelser/~/media/4206E200B65D47A8BCBA41EF9B3F1B31.ashx

39. Scott DJ, Dunnington GL. The new ACS/APDS Skills Curriculum: moving the learning curve out of the operating room. J Gastrointest Surg. Springer-Verlag; 2008 Feb;12(2):213–21.

40. Ghaderi I, Manji F, Park YS, Juul D, Ott M, Harris I, et al. Technical skills assessment toolbox: a review using the unitary framework of validity. Annals of Surgery. 2015 Feb;261(2):251–62.

41. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32(8):676–82.

42. Jang TB, Ruggeri W, Dyne P, Kaji AH. The Learning Curve of Resident Physicians Using Emergency Ultrasonography for Cholelithiasis and Cholecystitis. Acad Emerg Med. 2010 Nov;17(11):1247–52.

43. Frey Tirri B, Troeger C, Holzgreve W, Tercanli S. Quality management of nuchal translucency measurement in residents. Ultraschall Med. 2007 Oct;28(5):484–8.

44. Enriquez JL, Wu TS. An introduction to ultrasound equipment and knobology. Crit Care Clin. 2014 Jan;30(1):25–45–v.

45. Nicholls D, Sweet L, Hyett J. Psychomotor skills in medical ultrasound imaging: an analysis of the core skill set. J Ultrasound Med. American Institute of Ultrasound in Medicine; 2014 Aug;33(8):1349–52.

46. Magill R, Anderson D. Motor Learning and Control: Concepts and Applications. McGraw-Hill Higher Education; 2013.

47. Schmidt RA. A schema theory of discrete motor skill learning. Psychological Review. American Psychological Association; 1975 Jul 1;82(4):225–60.

48. Krupinski EA. The role of perception in imaging: past and future. Semin Nucl Med. 2011 Nov;41(6):392–400.

49. Drew T, Evans K, Võ ML-H, Jacobson FL, Wolfe JM. Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images? Radiographics. 2013 Jan;33(1):263–74.

50. Davis DP, Campbell CJ, Poste JC, Ma G. The association between operator confidence and accuracy of ultrasonography performed by novice emergency physicians. J Emerg Med. 2005 Oct;29(3):259–64.

51. Downing SM, Yudkowsky R. Assessment in Health Professions Education. Routledge; 2009. 1 p.

52. Miller GE. The assessment of clinical skills/competence/performance. Academic Medicine. 1990 Sep 1;65(9):S63.

53. Wulf G, Shea C, Lewthwaite R. Motor skill learning and performance: a review of influential factors. Medical Education. Wiley Online Library; 2010;44(1):75–84.

54. Guadagnoli MA, Lee TD. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. Journal of Motor Behavior. 2004 Jun;36(2):212–24.

55. Streiner DL, Geoffrey R Norman PD. Health Measurement Scales. Oxford University Press, USA; 2008. 1 p.

56. Pugh DM, Wood TJ, Boulet JR. Assessing Procedural Competence: Validity Considerations. Simul Healthc. 2015 Oct;10(5):288–94.

57. Shumway JM, Harden RM, Association for Medical Education in Europe. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Vol. 25, Medical Teacher. 2003. pp. 569–84.

58. Schuwirth LWT, van der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011;33(10):783–97.

59. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. Medical Education. 2002 Oct;36(10):972–8.

60. Ringsted C, Hodges B, Scherpbier A. "The research compass": an introduction to research in medical education: AMEE Guide no. 56. Vol. 33, Medical Teacher. 2011. pp. 695–709.

61. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. American Psychological Association; 1995 Sep 1;50(9):741–9.

62. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A practical guide to their development and use. OUP Oxford; 2014. 1 p.

63. Szasz P, Louridas M, Harris KA, Aggarwal R, Grantcharov TP. Assessing Technical Competence in Surgical Trainees: A Systematic Review. Annals of Surgery. 2015 Jun;261(6):1046–55.

64. Cook DA, Lineberry M. Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. Acad Med. 2016 Feb 2.

65. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 2014 Feb;36(2):97–110.

66. Yudkowsky R, Park YS, Lineberry M, Knox A, Ritter EM. Setting Mastery Learning Standards. Acad Med. 2015 Nov;90(11):1495–500.

67. Hays R. Standard setting. The Clinical Teacher. 2015 Aug;12(4):226–30.

68. Cizek GJ, Bunch MB. Standard setting. Sage Publications, Inc; 2007. 1 p.

69. Clauser BE, Clyman SG. A contrasting-groups approach to standard setting for performance assessments of clinical skills. Academic Medicine. 1994 Oct;69(10 Suppl):S42–4.

70. Livingston SA, Zieky MJ. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Educational Testing Service, Box 2885, Princeton, NJ 08541 ($7.50); 1982.

71. Tolsgaard MG, Ringsted C, Dreisler E, Klemmensen A, Loft A, Sorensen JL, et al. Reliable and valid assessment of ultrasound operator competence in obstetrics and gynecology. Ultrasound Obstet Gynecol. 2014 Apr;43(4):437–43.

72. Melchiors J, Todsen T, Nilsson P, Wennervaldt K, Charabi B, Bøttger M, et al. Preparing for Emergency: A Valid, Reliable Assessment Tool for Emergency Cricothyroidotomy Skills. Otolaryngol Head Neck Surg. 2014 Nov 10;152:260–5.

73. Konge L, Clementsen P, Larsen KR, Arendrup H, Buchwald C, Ringsted C. Establishing pass/fail criteria for bronchoscopy performance. Respiration. 2012;83(2):140–6.

74. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology. 2003 Oct;229(1):3–8.

75. Søreide K, Kørner H, Søreide JA. Diagnostic accuracy and receiver-operating characteristics curve analysis in surgical research and decision making. Annals of Surgery. 2011 Jan;253(1):27–34.

76. Dreyfus SE. The Five-Stage Model of Adult Skill Acquisition. bull sci technol soc. 2004 Jun 1;24(3):177–81.

77. Carraccio CL, Benson BJ, Nixon LJ, Derstine PL. From the educational bench to the clinical bedside: translating the Dreyfus developmental model to the learning of clinical skills. Acad Med. 2008 Aug;83(8):761–7.

78. Cheung JJH, Chen EW, Darani R, McCartney CJL, Dubrowski A, Awad IT. The Creation of an Objective Assessment Tool for Ultrasound-Guided Regional Anesthesia Using the Delphi Method. Regional Anesthesia and Pain Medicine. 2012 Apr 3;Publish Ahead of Print.

79. Chuan A, Graham PL, Wong DM, Barrington MJ, Auyong DB, Cameron AJD, et al. Design and validation of the Regional Anaesthesia Procedural Skills Assessment Tool. Anaesthesia. 2015 Dec;70(12):1401–11.

80. Wong DM, Watson MJ, Kluger R, Chuan A, Herrick MD, Ng I, et al. Evaluation of a task-specific checklist and global rating scale for ultrasound-guided regional anesthesia. Regional Anesthesia and Pain Medicine. 2014 Sep;39(5):399–408.

81. Konge L, Vilmann P, Clementsen P, Annema JT, Ringsted C. Reliable and valid assessment of competence in

endoscopic ultrasonography and fine-needle aspiration for mediastinal staging of non-small cell lung cancer. Endoscopy. 2012 Oct;44(10):928–33.

82. Davoudi M, Colt HG, Osann KE, Lamb CR, Mullon JJ. Endobronchial ultrasound skills and tasks assessment tool: assessing the validity evidence for a test of endobronchial ultrasound-guided transbronchial needle aspiration operator skill. Am J Respir Crit Care Med. 2012 Oct 15;186(8):773–9.

83. Konge L, Clementsen PF, Ringsted C, Minddal V, Larsen KR, Annema JT. Simulator training for endobronchial ultrasound: a randomised controlled trial. Eur Respir J. European Respiratory Society; 2015 Oct;46(4):1140–9.

84. Rice J, Crichlow A, Baker M, Regan L, Dodson A, Hsieh Y-H, et al. An Assessment Tool for the Placement of Ultrasound-Guided Peripheral Intravenous Access. J Grad Med Educ. 2016 May;8(2):202–7.

85. Clinkard D, Holden M, Ungi T, Messenger D, Davison C, Fichtinger G, et al. The development and validation of hand motion analysis to evaluate competency in central line catheterization. Acad Emerg Med. 2015 Feb;22(2):212–8.

86. Primdahl SC, Todsen T, Clemmesen L, Knudsen L, Weile J. Rating scale for the assessment of competence in ultrasound-guided peripheral vascular access - a Delphi Consensus Study. J Vasc Access. 2016;17(5):440-445.

87. Schmidt JN, Kendall J, Smalley C. Competency Assessment in Senior Emergency Medicine Residents for Core Ultrasound Skills. West J Emerg Med. 2015 Nov;16(6):923–6.

88. Bentley S, Mudan G, Strother C, Wong N. Are Live Ultrasound Models Replaceable? Traditional versus Simulated Education Module for FAST Exam. West J Emerg Med. 2015 Nov;16(6):818–22.

89. Ziesmann MT, Park J, Unger B, Kirkpatrick AW, Vergis A, Pham C, et al. Validation of hand motion analysis as an objective assessment tool for the Focused Assessment with Sonography for Trauma examination. The Journal of Trauma and Acute Care Surgery. 2015 Oct;79(4):631–7.

90. Ziesmann MT, Park J, Unger BJ, Kirkpatrick AW, Vergis A, Logsetty S, et al. Validation of the quality of ultrasound imaging and competence (QUICk) score as an objective assessment tool for the FAST examination. The Journal of Trauma and Acute Care Surgery. 2015 May;78(5):1008–13.

91. Sisley AC, Johnson SB, Erickson W, Fortune JB. Use of an Objective Structured Clinical Examination (OSCE) for the assessment of physician performance in the ultrasound evaluation of trauma. J Trauma. 1999 Oct;47(4):627–31.

92. Markowitz JE, Hwang JQ, Moore CL. Development and validation of a web-based assessment tool for the extended focused assessment with sonography in trauma examination. J Ultrasound Med. 2011 Mar;30(3):371–5.

93. Chung GKWK, Gyllenhammer RG, Baker EL, Savitsky E. Effects of simulation-based practice on focused assessment with sonography for trauma (FAST) window identification, acquisition, and diagnosis. Mil Med. 2013 Oct;178(10 Suppl):87–97.

94. Olszynski PA, Harris T, Renihan P, D'Eon M, Premkumar K. Ultrasound during Critical Care Simulation: A Randomized Crossover Study. CJEM. Cambridge University Press; 2015 Aug 26;:1–8.

95. Turner EE, Fox JC, Rosen M, Allen A, Rosen S, Anderson C. Implementation and assessment of a curriculum for bedside ultrasound training. J Ultrasound Med. American Institute of Ultrasound in Medicine; 2015 May;34(5):823–8.

96. Hofer M, Kamper L, Sadlo M, Sievers K, Heussen N. Evaluation of an OSCE assessment tool for abdominal ultrasound courses. Ultraschall Med. 2011 Apr;32(2):184–90.

97. Tolsgaard MG, Todsen T, Sorensen JL, Ringsted C, Lorentzen T, Ottesen B, et al. International multispecialty consensus on how to evaluate ultrasound competence: a delphi consensus survey. PLoS ONE. 2013;8(2):e57687.

98. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med. Springer-Verlag; 2005 Dec;20(12):1159–64.

99. Subhi Y, Todsen T, Konge L. An integrable, web-based solution for easy assessment of video-recorded performances. Adv Med Educ Pract. 2014;5:103–5.

100. Brennan RL. Generalizability Theory. Springer Verlag; 2001. 1 p.

101. Hojat M, Xu G. A visitor's guide to effect sizes–statistical significance versus practical (clinical) importance of research findings. Adv Health Sci Educ Theory Pract. Springer; 2004;9(3):241–9.

102. Guldbrand Nielsen D, Jensen SL, O'Neill L. Clinical assessment of transthoracic echocardiography skills: a generalizability study. BMC Med Educ. BioMed Central; 2015;15(1):9.

103. Crossley J, Russell J, Jolly B, Ricketts C, Roberts C, Schuwirth L, et al. 'I'm pickin“ up good regressions”: the governance of generalisability analyses. Medical Education. 2007 Oct;41(10):926–34.

104. Todsen T, Tolsgaard MG. Assessment of competence in FAST examination. The Journal of Trauma and Acute Care Surgery. 2016;80(2):353.

105. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013;4(2):627–35.

106. Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ringsted C. Using virtual-reality simulation to assess performance in endobronchial ultrasound. Respiration. 2013;86(1):59–65.

107. Damewood S, Jeanmonod D, Cadigan B. Comparison of a multimedia simulator to a human model for teaching FAST exam image interpretation and image acquisition. Acad Emerg Med. 2011 Apr;18(4):413–9.

108. Mandavia DP, Aragona J, Chan L, Chan D, Henderson SO. Ultrasound training for emergency physicians--a prospective study. Acad Emerg Med. 2000 Sep;7(9):1008–14.

109. Blehar DJ, Barton B, Gaspari RJ. Learning Curves in Emergency Ultrasound Education - Blehar - 2015 - Academic Emergency Medicine - Wiley Online Library. Academic Emergency …. 2015.

110. Knudson MM, Sisley AC. Training residents using simulation technology: experience with ultrasound for trauma. J Trauma. 2000 Apr;48(4):659–65.

111. Ali J, Rozycki GS, Campbell JP, Boulanger BR, Waddell JP, Gana TJ. Trauma ultrasound workshop improves physician detection of peritoneal and pericardial fluid. J Surg Res. 1996 Jun;63(1):275–9.

112. Keddis MT, Cullen MW, Reed DA, Halvorsen AJ, McDonald FS, Takahashi PY, et al. Effectiveness of an ultrasound training module for internal medicine residents. BMC Med Educ. BioMed Central; 2011;11(1):75.

113. Kotagal M, Quiroga E, Ruffatto BJ, Adedipe AA, Backlund BH, Nathan R, et al. Impact of point-of-care ultrasound training on surgical residents' confidence. J Surg Educ. 2015 Jul;72(4):e82–7.

114. Sekiguchi H, Bhagra A, Gajic O, Kashani KB. A general Critical Care Ultrasonography workshop: results of a novel Web-based learning program combined with simulation-based hands-on training. J Crit Care. 2012 Jul 2.

115. Mohammad A, Hefny AF, Abu-Zidan FM. Focused Assessment Sonography for Trauma (FAST) training: a systematic review. World J Surg. Springer US; 2014 May;38(5):1009–18.

116. Schuwirth LWT, van der Vleuten CPM. The use of clinical simulations in assessment. Medical Education. 2003 Nov;37 Suppl 1:65–71.

117. Gustafsson C, McNicholas A, Sondén A, Törngren S, Järnbert-Pettersson H, Lindelius A. Accuracy of Surgeon-Performed Ultrasound in Detecting Gallstones: A Validation Study. World J Surg. Springer International Publishing; 2016 Mar 2;:1–7.

118. Wyrick DL, Smith SD, Burford JM, Dassinger MS. Surgeon-performed ultrasound: accurate, reproducible, and more efficient. Pediatr Surg Int. Springer Berlin Heidelberg; 2015 Aug 12;31(12):1–4.

119. Hamer PW, Aspinall SR, Malycha PL. Clinician-performed ultrasound in assessing potentially malignant thyroid nodules. ANZ J Surg. 2014 May;84(5):376–9.

120. Badran K, Jani P, Berman L. Otolaryngologist-performed head and neck ultrasound: outcomes and challenges in learning the technique. J Laryngol Otol. 2014 May;128(5):447–53.

121. McCarter FD, Luchette FA, Molloy M, Hurst JM, Davis K, Johannigman JA, et al. Institutional and individual learning curves for focused abdominal ultrasound for trauma: cumulative sum analysis. Annals of Surgery. 2000 May;231(5):689–700.

122. Brighton B, Bhandari M, Tornetta P, Felson DT. Hierarchy of evidence: from case reports to randomized controlled trials. Clinical Orthopaedics and Related Research. 2003 Aug;413(413):19–24.

123. Cook DA, Beckman TJ. Reflections on experimental research in medical education - Springer. Adv in Health Sci Educ. 2010.

124. Cook DA, Bordage G, Schmidt HG. Description, justification and clarification: a framework for classifying the purposes of research in medical education - Cook - 2008 - Medical Education - Wiley Online Library. Medical Education. 2008.

125. Davis D, O'Brien MAT, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of Formal Continuing Medical Education: Do Conferences, Workshops, Rounds, and Other Traditional Continuing Education Activities Change Physician Behavior or Health Care Outcomes? JAMA [Internet]. American Medical Association; 1999 Sep 1;282(9):867–74. Available from: http://archneur.jamanetwork.com/article.aspx?articleid=191423

126. Todsen T, Henriksen MV, Kromann CB, Konge L, Eldrup J, Ringsted C. Short- and long-term transfer of urethral catheterization skills from simulation training to performance on patients. BMC Med Educ. BioMed Central Ltd; 2013;13(1):29.

127. AlEassa EM, Ziesmann MT, Kirkpatrick AW, Wurster CL, Gillman LM. Point of care ultrasonography use and training among trauma providers across Canada. Can J Surg. 2016 Feb;59(1):6–8.

128. Ringsted C, Skaarup AM, Henriksen AH, Davis D. Person-task-context: a model for designing curriculum and in-training assessment in postgraduate education. Med Teach. 2006 Feb;28(1):70–6.

129. Hillingsø JG, Bo Svendsen L, Bachmann Nielsen M. Focused bedside ultrasonography by clinicians: Experiences with a basic introductory course. Scand J Gastroenterol. Informa UK Ltd UK; 2008 Jan;43(2):229–33.

130. Nielsen DG, Gotzsche O, Sonne O, Eika B. The relationship between immediate relevant basic science knowledge and clinical knowledge: physiology knowledge and transthoracic echocardiography image interpretation. Adv in Health Sci Educ. Springer Netherlands; 2012 Oct;17(4):501–13.

131. Østergaard ML, Ewertsen C, Konge L, Albrecht-Beste E, Bachmann Nielsen M. Simulation-Based Abdominal Ultrasound Training - A Systematic Review. Ultraschall Med. 2016 Feb 16.

132. Tolsgaard MG, Ringsted C, Rosthøj S, Nørgaard L, Møller L, Freiesleben NLC, et al. The Effects of Simulation-based Transvaginal Ultrasound Training on Quality and Efficiency of Care: A Multicenter Single-blind Randomized Trial. Annals of Surgery. 2016 Jan 25.

133. Foss KT, Subhi Y, Aagaard R, Bessmann EL, Bøtker MT, Graumann O, et al. Developing an emergency ultrasound app - a collaborative project between clinicians from different universities. Scand J Trauma Resusc Emerg Med. BioMed Central Ltd; 2015;23(1):47.

134. Melchiors J, Todsen T, Nilsson P, Kohl AP, Bøttger M, Charabi B, et al. Self-directed simulation-based training of emergency cricothyroidotomy: a route to lifesaving skills. Eur Arch Otorhinolaryngol; 2016;273(12):1–6.

135. Gazelle GS, Kessler L, Lee DW, McGinn T, Menzin J, Neumann PJ, et al. A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. Radiology. 2011 Dec;261(3):692–8.

136. Raguin T, Schneegans O, Rodier J-F, Volkmar P-P, Sauleau E, Debry C, et al. Value of fine-needle aspiration in evaluating large thyroid nodules. Head Neck. 2016 Jun 14.

137. Carneiro-Pla D, Amin S. Comparison Between Preconsultation Ultrasonography and Office Surgeon-Performed Ultrasound in Patients with Thyroid Cancer. World J Surg. 2013 Oct 19.

138. Ahn D, Kim H, Sohn JH, Choi JH, Na KJ. Surgeon-performed ultrasound-guided fine-needle aspiration cytology of head and neck mass lesions: sampling adequacy and diagnostic accuracy. Ann Surg Oncol. Springer US; 2015 Apr;22(4):1360–5.

139. Fernandes VT, Magarey MJR, Kamdar DP, Freeman JL. Surgeon performed ultrasound-guided fine-needle as- pirates of the thyroid: 1067 biopsies and learning curve in a teaching center. Head Neck. 2016 Apr;38 Suppl 1(S1):E1281–4.

140. Coltrera MD. Clinician-performed thyroid ultrasound. Otolaryngol Clin North Am. 2014 Aug;47(4):491–507.

141. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach. 2012;34(11):960–92.

142. Portney LG, Watkins MP. Foundations of Clinical Re- search. Prentice Hall; 2009. 1 p.

143. Greiner M, Pfeiffer D, Smith RD. Principles and practi- cal application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med. 2000 May 30;45(1-2):23–41.

144. Orloff LA, Randolph GW. Preoperative Imaging for Thyroid Cancer: Beyond Ultrasonography. JAMA Oto- laryngol Head Neck Surg. 2016 Mar 17.

145. Mazzaglia PJ. Surgeon-performed ultrasound in pa- tients referred for thyroid disease improves patient care by minimizing performance of unnecessary pro- cedures and optimizing surgical treatment. World J Surg. 2010 Jun;34(6):1164–70.

146. Hwang HS, Orloff LA. Efficacy of preoperative neck ultrasound in the detection of cervical lymph node me- tastasis from thyroid cancer. Laryngoscope. 2011 Mar;121(3):487–91.

147. Akbar NA, Bodenner DL, Kim LT, Suen JY, Kokoska MS. Considerations in incorporating office-based ultra- sound of the head and neck. Otolaryngol Head Neck Surg. 2006 Dec;135(6):884–8.

148. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991 Apr;11(2):88–94.